Orchestrating AI Agents – Model Context Protocol (MCP)

Hafsa Nawaz
04/26/2025

Link to demo: https://youtu.be/aWyM9jGCBZE

Introduction 1
How MCP Facilitated Communication 1
Model Parameters and Technical Choices 1
Strengths Observed 2
Limitations and Observed Challenges 2
Emergent Behaviors 3
Recommendations for Future Improvements 3

Introduction

This project implemented a multi-agent debate system orchestrated using the Model Context Protocol (MCP) framework. Two AI agents, Agent Pro and Agent Con, engaged in a structured debate on real-world topics with minimal human intervention. The agents were powered by Groq's hosted version of the Llama 3 8B model, accessed via API. The system emphasized message flow, role persistence, and autonomous dialogue generation.

How MCP Facilitated Communication

The MCP structure standardized interaction between agents by enforcing consistent message formats: each message included a sender and content. The orchestrator seeded the conversation with a debate topic, and each agent generated its contribution based on the evolving context. By passing structured Message objects back and forth, the agents could read prior messages and respond meaningfully without manual edits. This created a natural multi-turn dialogue flow aligned with real-world debate protocols.

Model Parameters and Technical Choices

- **Temperature = 0.7**:

    - Setting the temperature to 0.7 aimed to balance creativity and consistency. A lower temperature (e.g., 0.2) would make the model overly deterministic and repetitive, while a higher temperature (e.g., 1.0) could lead to unpredictable or incoherent arguments.
    - At 0.7, the agents produced formal, diverse yet logically grounded arguments, enhancing debate quality without sacrificing structure.
- **Max Tokens = 512**:
    - This limit was chosen to ensure concise yet complete arguments.
    - It prevented overly long, rambling answers that could derail the debate's pacing.
- **Groq API Usage**:
    - Groq's Llama 3 model was selected for its speed and reliability, allowing rapid multi-turn exchanges.
    - This minimized wait times between agent turns and simulated a more realistic back-and-forth debate dynamic.

Strengths Observed

- **Role Consistency**:
    - Agents consistently defended their assigned perspectives (pro/con) across turns without drifting off-topic.
- **Autonomous Continuity**:
    - Each agent not only responded to the immediate prior argument but also maintained long-term thematic consistency (e.g., Agent Pro consistently emphasized ethics and safety, while Agent Con focused on innovation and freedom).
- **Formal Tone Emergence**:
    - Despite no explicit instruction for debate etiquette, agents naturally adopted formal opening and closing statements.
- **Speed and Smoothness**:
    - Using Groq's infrastructure allowed a natural pacing for the conversation without disruptive lags.

Limitations and Observed Challenges

- **Surface-Level Rebuttals**:
    - Occasionally, agents rebutted general ideas rather than directly dismantling the opponent's specific strongest points.
- **Phrase Repetition**:

    - Standardized formal phrases ("Thank you, esteemed judges...") appeared excessively across multiple turns, potentially reducing variation.
- **Predictable Structure**:
    - Since the debate strictly followed an Opening → Rebuttal → Closing structure, there was limited adaptability if a new line of attack emerged during rebuttals.
- **Memory Overload Risk**:
    - The agents relied on full message history context, which could potentially lead to degraded focus if debates were extended to many turns without a memory window.

Emergent Behaviors

- **Polite Formality**:
    - Both agents maintained high decorum even under rebuttal, mimicking human academic debate styles.
- **Thematic Persistence**:
    - Agents anchored their arguments to core values (ethics vs innovation) without needing explicit memory engineering.
- **Self-Correction Tendencies**:
    - Occasionally, agents would preemptively address potential counter-arguments from the opponent, showcasing anticipatory behavior.

Recommendations for Future Improvements

- **Dynamic Argument Scoring**:
    - Implement a lightweight scoring system to prioritize rebutting the most impactful points raised by the opponent rather than treating all arguments equally.
- **Memory Trimming**:
    - Introduce a sliding window memory (e.g., last 3–4 messages) to keep agent focus sharp over longer debates.
- **Softer Formality**:
    - Introduce randomization or prompt variation to avoid robotic repetition of standard phrases like "esteemed judges" every turn.
- **Crossfire Section**:
    - Design an active "cross-examination" round where agents directly question each other's arguments before closing statements.